The Repatterning of Eukaryotic Genomes by Random Genetic Drift

Michael Lynch,[1] Louis-Marie Bobay,[2] Francesco Catania,[3] Jean-François Gout,[1] and Mina Rho[4]

[1]Department of Biology and [4]Department of Computer Science, Indiana University, Bloomington, Indiana 47408; email: milynch@indiana.edu

[2]Microbial Evolutionary Genomics, Institut Pasteur, CNRS, URA2171, F-75724 Paris, France

[3]Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom

Annu. Rev. Genomics Hum. Genet. 2011. 12:347–66

First published online as a Review in Advance on July 13, 2011

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi: 10.1146/annurev-genom-082410-101412

1527-8204/11/0922-0347$20.00

Keywords

complexity, genome evolution, mutation, protein evolution, recombination

Abstract

Recent observations on rates of mutation, recombination, and random genetic drift highlight the dramatic ways in which fundamental evolutionary processes vary across the divide between unicellular microbes and multicellular eukaryotes. Moreover, population-genetic theory suggests that the range of variation in these parameters is sufficient to explain the evolutionary diversification of many aspects of genome size and gene structure found among phylogenetic lineages. Most notably, large eukaryotic organisms that experience elevated magnitudes of random genetic drift are susceptible to the passive accumulation of mutationally hazardous DNA that would otherwise be eliminated by efficient selection. Substantial evidence also suggests that variation in the population-genetic environment influences patterns of protein evolution, with the emergence of certain kinds of amino-acid substitutions and protein-protein complexes only being possible in populations with relatively small effective sizes. These observations imply that the ultimate origins of many of the major genomic and proteomic disparities between prokaryotes and eukaryotes and among eukaryotic lineages have been molded as much by intrinsic variation in the genetic and cellular features of species as by external ecological forces.

INTRODUCTION

It is generally acknowledged that the biological world was entirely prokaryotic 3 billion years ago, and that by ~2.5 billion years ago, a key lineage had emerged that eventually gave rise to all of today's eukaryotes. Although prokaryotes are the evolutionary cradle of metabolic diversity and still dominate the earth numerically, the emergence of eukaryotes initiated an enormous radiation of morphological diversity. The factors responsible for this diversification and the degree to which natural selection played a role remain unclear.

Based on the attributes shared across the entire eukaryotic domain, we can be reasonably certain that the ancestral eukaryotic cell harbored a complex genome and a complex internal structure (47, 65, 83), and it is tempting to further assume that the emergence of internal membrane-bound structures was a necessary precursor to the evolution of diverse external morphologies (50). However, although cellular features comprise the physical substrate upon which natural selection operates, intrinsic processes operating at the population-genetic level dictate the types of paths that are open or closed to evolutionary exploration within different phylogenetic lineages. Here, we review the evidence that alterations in the population-genetic environment played a central and possibly definitive role in establishing the unique types of evolutionary trajectories taken by various eukaryotic lineages at the genomic and proteomic levels.

As the subject material is broad, we will restrict our attention to three fundamental issues. First, we will examine the general phylogenetic patterns of the three main nonadaptive features of the population-genetic environment—random genetic drift, recombination, and mutation—as the relative powers of these forces define the types of evolutionary changes that are possible in various contexts. Second, we will review the broad set of observations on eukaryotic genome structure that have emerged via the field of comparative genomics. With considerably more data than were available in the past, this overview will clearly establish the general boundaries of the overall genome-architectural landscape within which eukaryotic lineages have wandered over evolutionary time. Third, having established the central role that random genetic drift and mutation have played in the diversification of genome structure, we will move to the next rung of the ladder in biological organization, the nature of the proteome, providing suggestive arguments that alterations in the nonadaptive forces of evolution in the eukaryotic domain are of sufficient magnitude to influence the ways in which protein evolution proceeds. New advances in population-genetic theory in this area provide a potential resource for understanding how complex cellular adaptations may have evolved in the ancestral eukaryote.

THE POPULATION-GENETIC ENVIRONMENT

Although natural selection plays an essential role in molding organismal diversity in ways that ensure population survival, the stochastic nature of the processes of random genetic drift, recombination, and mutation makes it impossible to precisely predict the genomic or phenotypic responses that will be elicited by any specific selective challenge. However, two things are clear. First, evolution follows the dictates of Darwin's "descent with modification"—natural selection operates on standing variation, with the new variants that arise by mutation and recombination being defined by the preexisting resources. Second, the ability of natural selection to promote beneficial mutations and eradicate deleterious mutations depends on the intensity of selection at the gene level relative to the power of random genetic drift. If the magnitude of drift exceeds the power of selection, adaptations that would otherwise go forward cannot be actively promoted, whereas degenerative mutations with sufficiently mild effects will accumulate. Although the latter effect can lead to extinction of a sufficiently small population (73, 74), it can also promote novel paths of evolution as initially deleterious

mutations that passively emerge at the DNA level are secondarily modified into new adaptive forms. Here, we briefly review how the magnitudes of the three major features of the population-genetic environment scale across the tree of life.

Random Genetic Drift

Random sampling of the gene pool from generation to generation is a ubiquitous source of evolutionary stochasticity. The magnitude of drift is generally defined by the inverse of the effective number of gametes sampled per generation—$2N_e$ in a diploid species, where N_e is the effective population size (15). Although N_e is generally expected to increase with the absolute number of reproductive adults in a species (N), many additional factors influence patterns of gene transmission across generations. Most aspects of population structure, such as uneven sex ratios, variation in family size, nonrandom mating, and localized inbreeding, result in nonrandom representation of genomes across generations, guaranteeing that $N_e < N$—i.e., that fluctuations of allele frequencies across generations will be much larger than expected were gametes to be sampled equally from N parents.

However, of at least equal importance is the fact that the physical structure of the genome ensures that N_e will be further depressed below the expectation based on gamete sampling. This is because the fates of nucleotides at a specific genomic site are determined by selection operating not just on that site, but also on all linked sites under selection. As a consequence, the direct effects of some deleterious mutations can be partially masked by fortuitous linkage to beneficial mutations and also by the simultaneous presence of competing deleterious mutations in other individuals. In the extreme case of an obligately asexual species, N_e is not much more than the number of individuals in the highest fitness class of the population, as essentially all other individuals represent the "living dead" who will leave no long-term descendants (29). Likewise, the ability of selection to promote new adaptive mutations is reduced by linkage, as favorable (nonallelic) mutations arising on homologous chromosomes cannot be simultaneously fixed unless recombination occurs between the two sites. These interference effects from linkage are expected to increase with increasing N, as a larger number of mutational targets enhances the likelihood of simultaneous segregation of multiple mutations (28). Consequently, the relationship between N_e and N is almost certainly nonlinear, with the response of N_e to N becoming shallower (and perhaps even leveling off) at very large N, as the quantitative consequences of linkage (draft) overtake the magnitude of drift associated with gamete sampling. The net effect of these complexities is that $1/(2N_e)$ should be viewed simply as a composite summary measure of the long-term consequences of all sources of transmission stochasticity, including those caused by linkage, mating-system variation, etc.

In principle, N_e can be estimated directly by monitoring the variance of allele-frequency change across generations, as this has an expected value $p(1 - p)/(2N_e)$, where p is the initial allele frequency (88, 106). However, as the expected change in allele frequency is extremely small unless N_e is tiny, this approach is notoriously difficult to apply because errors in estimating p will overwhelm the true change unless the sample size is enormous. As a consequence, most attempts to estimate N_e have taken a circuitous route, the most popular being indirect inference from observations on levels of within-population variation at nucleotide sites assumed to be neutral. The logic underlying this approach is that if u is the rate of base-substitution mutation per generation for the sites under consideration, and N_e is roughly constant, an equilibrium level of variation will be reached at which the input via mutation, $2u$, is matched by the fractional loss via drift/draft, $1/(2N_e)$. At this point, the level of nucleotide heterozygosity is approximately equal to the ratio of these two forces, $4N_eu$ for a diploid ($2N_eu$ for haploids), provided the observed heterozygosity $\ll 1$, which is almost always the case. The obvious limitation of this estimator is that it is

not simply a function of N_e, but also depends on u. On the other hand, as will be noted below, a central parameter in many areas of genome evolution is the ratio of the magnitudes of mutation and drift, $u/[1/(2N_e)] = 2N_eu$, so this composite population-genetic estimator is in fact of great utility (65).

After factoring out the contribution from mutation (below), standing levels of variation at silent sites imply $N_e \cong 10^5$ for vertebrates, $\sim 10^6$ for invertebrates and land plants, $\sim 10^7$ for unicellular eukaryotes (and fungi), and $>10^8$ for free-living prokaryotes (66). Although crude, these estimates imply that the power of drift is substantially elevated in eukaryotes—e.g., at least three orders of magnitude in large multicellular species relative to prokaryotes. It is also clear that the genetic effective sizes of populations are generally very far below actual population sizes.

Mutation

Although mutation rates have historically been estimated by indirect methods such as reporter constructs and phylogenetic analysis, the recent application of high-throughput sequencing methods to long-term mutation-accumulation lines has yielded highly refined genome-wide estimates of the mutation rate in several model systems (66). The results indicate a clear increase in the mutation rate with increasing genome size and organism complexity, from an average $<10^{-9}$ base substitutions per site per generation in prokaryotes to $>10^{-8}$ in mammals, with invertebrates and land plants exhibiting intermediate values. Results from in vitro assays of replicative polymerases and the mismatch-repair enzymes suggest that these modifications are due, at least in part, to variation in the efficiency of the replication/repair machinery (69), although multicellular species may also be capable of minimizing germline mutation rates per cell division by maintaining a relatively nonmutagenic environment in such cell lineages.

When combined with the above-noted estimates of standing variation on silent-site heterozygosity, these direct observations further imply an inverse relationship between the mutation rate and N_e in accordance with the drift-barrier hypothesis for mutation-rate evolution (66, 69). The mechanistic underpinnings of this hypothesis are derived from two generalities. First, owing to the predominance of deleterious mutations, selection generally operates to reduce the mutation rate. Second, selection on the mutation rate is a second-order effect, in that the magnitude of selection against a weak mutator allele is a function of the excess number of deleterious mutations that it promotes and remains in linkage disequilibrium with. In sexually reproducing species, the association with unlinked mutations will be eliminated in just two generations on average, whereas mutators in asexual populations acquire a higher load, as associations with mutations are retained indefinitely unless removed by the slower action of selection. The net consequence of these effects is that the selective advantage of an antimutator allele is generally on the order of the reduction in the genome-wide deleterious-mutation rate in asexual populations and s times that in sexual populations (18, 41, 45), where s is the average selective disadvantage of a deleterious mutation [which is typically on the order of 0.001 to 0.01 (78)].

From these observations, it can be concluded that because genome-wide deleterious-mutation rates are typically on the order of 0.01 to 1.0 (78) and are undoubtedly influenced by many dozens of loci, most single-amino-acid substitutions in replication/repair loci will have very small selective consequences. Once the mutation rate has been driven down to the level at which the improvement of fitness by antimutators is typically less than the magnitude of drift, $\sim 1/(2N_e)$, the lower bound to the mutation rate has been reached. Thus, the increase in the per-generation mutation rate in eukaryotes relative to prokaryotes, and in multicellular species in particular, is likely to be a simple pathological consequence of a reduction in N_e, and not an outcome of selection to improve the long-term evolvability of such taxa.

Recombination

As in the case of the mutation rate, there is a strong association between recombination rates and effective population sizes, although in this case the causal effect is indirect. As will be discussed below, eukaryotic genome sizes substantially increase in response to declining N_e, but such expansions are generally a consequence of increases in chromosome size rather than chromosome number. Thus, as a consequence of a nearly invariant feature of meiosis—the occurrence of approximately one crossover event per chromosome arm per meiotic event—the average rate of recombination per unit of physical distance exhibits a strong inverse relationship with genome size (**Figure 1**). Most of the remaining noise around this relationship is due to variation in haploid chromosome number (C), with the average amount of recombination per unit of physical distance on chromosomes being closely approximated by $2C/G$ for sexually reproducing eukaryotes, where G is the haploid genome size.

Although this pronounced pattern broadly defines the average levels of recombinational activity across the eukaryotic tree of life, it is also true that considerable variation exists in local rates of recombination within chromosomes and within and among individuals (20, 90). Nonetheless, although recombination hot spots are common (7, 38), they appear to be transient phylogenetically, and the evidence suggests that recombination-rate variation may be largely a function of neutral processes (19). Indeed, recombination hot spots are expected to be self-destructive, as they are typically converted by their less recombinogenic allelic partners during gene conversion events (86).

Thus, although considerable theoretical effort has been addressed toward understanding the role of adaptation in promoting recombination-rate modifiers (6, 33, 43), the bulk of the evidence suggests that recombination rates per meiotic event are not fine-tuned by natural selection. If anything, given the conserved deployment of just one to two crossover events per chromosome across the entire eukaryotic domain, selection appears to minimize genetic exchange within chromosomes, perhaps to minimize the damage that can accompany double-strand breaks. The simplest ways to modify global recombination rates are then to either alter the numbers of chromosomes (for which there is no phylogenetic pattern) or restrict the frequency of meiosis (in species capable of alternating episodes of sexual and asexual propagation).

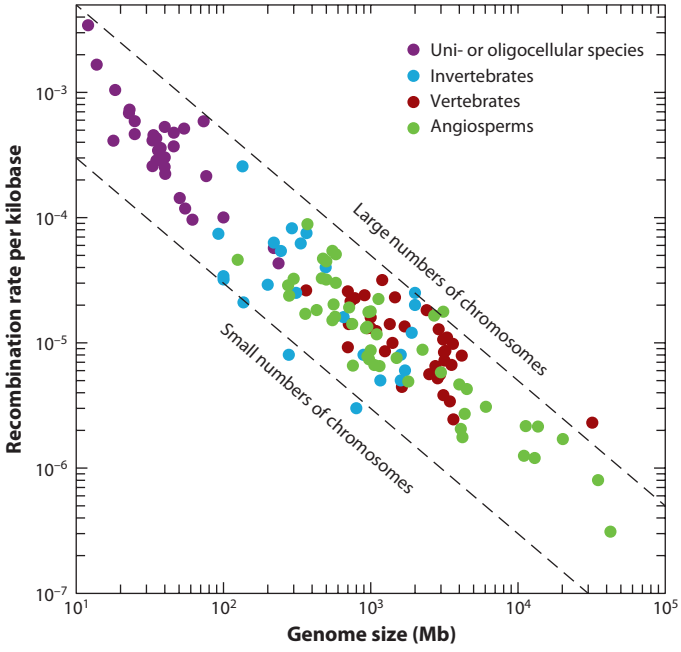

Figure 1

A compilation of estimates of the average amount of recombination per unit of physical distance in eukaryotic genomes, derived from 137 meiotic genetic maps. The diagonal lines have slopes of −1.

The net result of the per-site recombination rate declining and the mutation rate increasing with genome size is that the ratio of rates of recombination to mutation per nucleotide site increases from ∼1.0 in multicellular species to >100 in unicellular eukaryotes (65). However, the quantitative consequences of these differences in recombinational activity for genome evolution are not entirely clear. Because the large genomes of multicellular eukaryotes often contain >90% noncoding DNA with unknown (and in many cases,

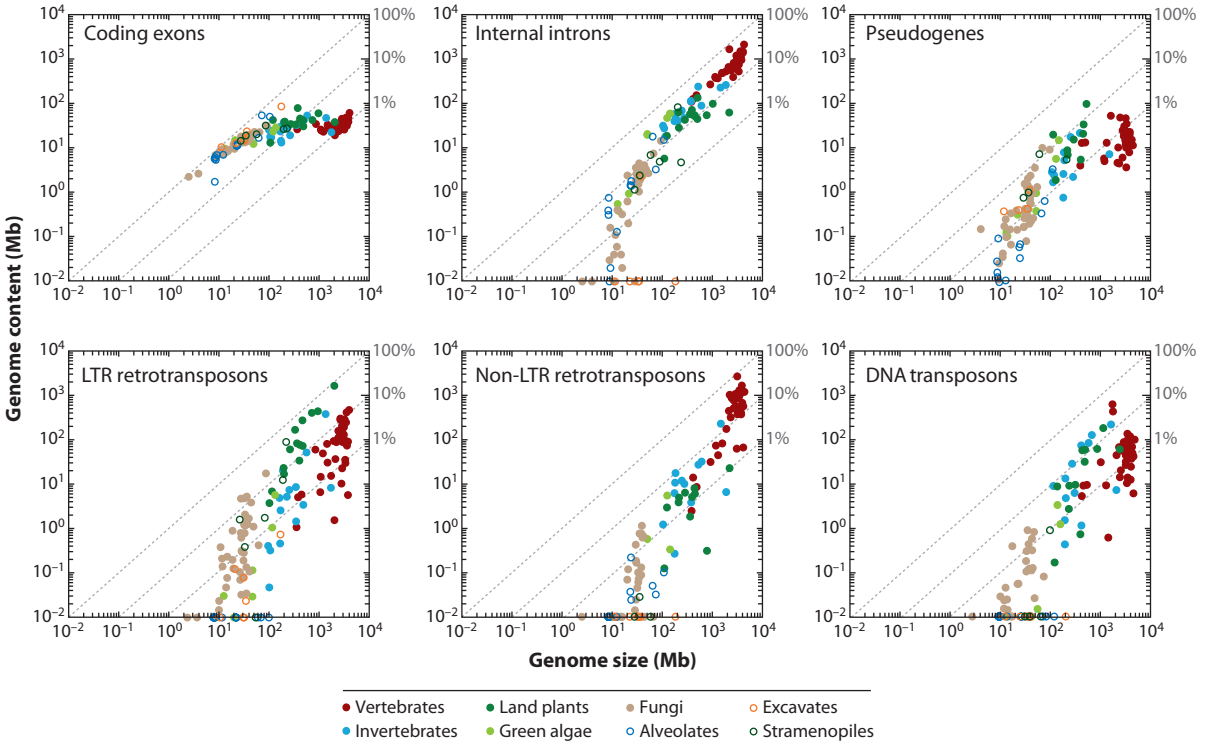

Figure 2

The scaling of genome content with genome size across ~150 eukaryotes. The full data set is available from the authors upon request. Note that the pool of intronic DNA can contain mobile elements and/or pseudogenes. Abbreviation: LTR, long terminal repeat.

probably nonexistent) functions, it is likely that the disparity in rates of recombination between selected sites is less than that suggested above, owing to the greater dispersion of such sites in species with greater abundance of spacer DNA.

GENOME ARCHITECTURE AND GENE-STRUCTURAL COMPLEXITY

The observations outlined in the preceding paragraphs clearly indicate that the evolution of eukaryotes was accompanied by the emergence of major differences in the population-genetic environment, with species in multicellular lineages exhibiting extreme lows for the power of recombination and extreme highs in the rates of mutation and magnitude of random genetic drift. Together, these changes produce a synergistic combination of conditions that increase the likelihood of mildly deleterious mutation accumulation while reducing the ability of selection to promote beneficial mutations with small advantages. The original argument that the consequences of such a syndrome are reflected on a genome-wide scale in a fairly regular way across the entire tree of life was based on a relatively small number of taxa (72), but it is now possible to expand this earlier survey to ~150 eukaryotic species (**Figure 2**).

As genome sizes increase from ~1 Mb in the smallest eukaryotes to >10 Gb in the largest, the most pronounced change is a progressive increase in the contribution of noncoding DNA. This pattern is common to all forms of noncoding sequence, from the three main classes of mobile elements to introns to pseudogene-associated sequences. Variation exists in the average positioning of major phylogenetic groups on an axis of genome size, but there is

an overall continuity of scaling between groups such that the content of a large fungal genome approximates that of an invertebrate genome of comparable size, etc. This general syndrome of genomic bloating, operating in parallel across eukaryotic lineages, is thought to be a consequence of the reduced efficiency of selection opposing the accumulation of excess DNA in taxa with reduced N_e and a mutational bias toward insertions of large segments of DNA (62, 65).

The underlying premise of this hypothesis is that essentially all forms of excess DNA are mutationally hazardous. For example, the addition of every intron to a gene imposes a constraint on ~30 nucleotides of gene sequence required for proper recognition by the spliceosome (67). The expanded 5′ untranslated regions of eukaryotic genes serve as mutational substrate for the origin of premature translation-initiation codons, which typically result in defective transcripts (77). In addition, although completely nonfunctional intergenic DNA cannot suffer from a loss-of-function mutation, it can incur mutations that cause adjacent genes to turn on at inappropriate times or places (27, 32).

As will be described below, the mutational-hazard hypothesis provides a unifying explanatory framework for a broad array of previously disconnected observations on genome size and gene structure. A few other hypotheses have been proposed as explanations for variation in genome size, but these have no obvious connection to the issue of gene-structural evolution, and have other limitations as well. For example, Cavalier-Smith (13, 14) has been a persistent advocate of the view that genome size (independent of gene content) is a quantitative trait selected upon for its influence on a variety of interrelated cellular features, including the size of the nuclear envelope and the flow of transcripts to the cytoplasm. However, as noted in **Figure 2**, the primary mechanism of genome-size expansion in eukaryotes is the proliferation of diverse families of mobile elements, all of which serve as major sources of deleterious mutation, and it is unclear how often the major mutational burden imposed by such elements (resulting from the inactivation of host genes into which they hop) can be offset by any cellular advantages of bulk DNA. One situation in which such a barrier might be at least transiently surmounted is whole-genome duplication.

From a quite different perspective, Hessen et al. (34) have suggested that patterns of nutrient limitation in various lineages select for alternative strategies for investing in DNA versus RNA, most notably the reallocation of nitrogen and phosphorus from the genome to ribosomes in rapidly growing species. However, as noted by Vieira-Silva et al. (105; see also 35), there is no relationship between genome size and cell growth rates across a wide range of prokaryotes, and through increases in the numbers of origins of replication, large genomes can come to replicate just as rapidly as small genomes (65). Thus, there is no evidence that genome sizes are constrained by energetic or time constraints (with viruses being a possible exception).

One of the primary predictions of the mutational-hazard hypothesis is that the susceptibility of a genome to the accumulation of excess DNA is a function of the ratio of the power of mutation to drift (65). Consider a modification to a gene's structure that increases the susceptibility to degenerative mutations. If such an embellishment increases the mutational target size by n nucleotides (e.g., as noted above, $n \approx 30$ in the case of introns), and the mutation rate per site is u, a new allele carrying such a change will experience an excess rate of degradation to defective alleles equal to nu. This differential susceptibility to mutation operates exactly like selection, in that the probability of fixation of a structurally modified allele (assuming no change in the protein sequence or other functional aspects) is defined by the standard diffusion approximation of Kimura (44), $2s/(1 - e^{-4N_e s})$, with the selective disadvantage being set equal to $s = -nu$ (61). The quantity $4N_e s$ is equivalent to the ratio of the relative strength of selection, s, and drift, $1/(2N_e)$. If $|4N_e s| \ll 1$, drift dominates the dynamics of allele-frequency change, and the probability of fixation of a deleterious allele is closely

approximated by the initial frequency $1/(2N_e)$, whereas if $|4N_e s| \gg 1$, selection dominates, and the probability of fixation of a deleterious allele asymptotically approaches zero.

Thus, the likelihood of a mutationally hazardous modification fixing depends critically on whether the composite quantity $4N_e nu$ is greater or less than 1. Rewriting the latter condition as $4N_e u < 1/n$, to denote the point where the efficiency of selection is compromised enough that a harmful allele (or gene region) can readily drift to fixation, clarifies a central point of the mutational-hazard hypothesis. Although it is commonly implied that the key parameter underlying this hypothesis is the effective population size (e.g., 31, 100, 112), this is not strictly true. Rather, the likelihood of success of a mutationally hazardous gene modification is the ratio of the power of mutation to drift, $2N_e u$. One of the more dramatic sources of support for this point derives from observations on organelle evolution (75). Although the nuclear genomes of land plants and animals have independently arrived at quite similar architectural states (**Figure 2**), these lineages have evolved enormous differences in the structure of organelle genomes—those in land plants are bloated with intergenic DNA and introns, whereas those in animals are extremely compact, to the point that the translation-initiation and translation-termination codons of adjacent genes often overlap. These dramatic disparities can be explained by the fact that mutation rates in plant organelles are typically ~100 times lower than those in animal organelles, whereas the nuclear mutation rates and average values of N_e are roughly comparable across both lineages (65).

The plausibility of the mutational-hazard hypothesis derives from the fact that the mutational costs of various additions to eukaryotic genes are generally weak enough that such modifications are vulnerable to passive accumulation in species with sufficiently small N_e. Consider, for example, the consequences of adding an intron that imposes constraints on the sequences at ~30 key nucleotide sites (67). Given that the human mutation rate per nucleotide site is $\sim 1.3 \times 10^{-8}$ per generation (67), the selective disadvantage (*nu*) of a newly arisen intron-containing allele in the human lineage is then $\sim 4 \times 10^{-7}$. Because the basic splicing machinery is fairly conserved across all eukaryotes, this cost is likely to be quite general over all lineages (with the caveat that mutation-rate differences must be accounted for, yielding a lower value of *nu* for unicellular species). Thus, as a first-order approximation, N_e in excess of 10^6 is required for selection to be effective at preventing the fixation of a new intron-containing allele in a multicellular eukaryote (assuming no additional costs associated with the intron). As noted above, the effective population sizes of most land plants and animals are near or well below this threshold, whereas estimates of N_e for unicellular eukaryotes are in the vicinity of the threshold and may often exceed it. Thus, the association of the position of the threshold behavior for average intron investment with unicellular lineages (**Figure 2**) is consistent with theoretical expectations.

The scatter around the general gradients observed in **Figure 2** is entirely expected, as species near the threshold can be expected to wander in both directions over evolutionary time, and genome size cannot respond instantaneously to such changes. For example, because many mechanisms exist for both the gain and loss of excess DNA, the mutational-hazard hypothesis implies that if N_e were to expand after a sufficiently prolonged earlier pattern of lower N_e, all aspects of genome architecture should undergo a gradual modification in response to the change in the population-genetic environment. A most pronounced example of such behavior is observed in the age-distributional patterns for a variety of insertions in mammalian genomes (97). Long terminal repeat (LTR) retrotransposons can be aged from the divergence of their terminal sequences, which are identical at the time of birth of a new element, and the times of origin of inclusions such as processed pseudogenes and fragments of nuclear insertions of mitochondrial DNA (numts) can be estimated by reference to their native gene sources. In a wide variety of

invertebrates, land plants, and unicellular eukaryotes, the age distributions of such insertions exhibit negative-exponential forms, consistent with a long-term steady-state birth/death process (65). However, mammalian genomes exhibit dramatic age-distributional bulges for a wide variety of insertion types, which can only be explained by recent increases in loss rates and/or recent decreases in insertion rates of such elements.

There are three remarkable features of the mammalian data. First, the peaks of the age distributions of inserts occur at roughly comparable times, ~35–60 Mya, across a wide array of mammalian orders. Thus, because the orders of mammals began to diverge ~100 Mya (113), these changes represent a dramatic example of parallel genomic evolution. Second, the dates of the recent declines in the age distributions of inserts vary among different types of inserts, with peaks appearing ~60 Mya for pseudogenes, 30 Mya for LTR retrotransposons, and 20 Mya for numts. Third, extrapolation from the age distributions of inserts implies that mammalian genome sizes prior to 60 Mya were approximately double those in modern-day species, and that today's genomes are still in a contraction phase—i.e., have not reached equilibrium.

Our interpretation of these results is that the extinction of the dinosaurs at the K-T boundary (~65 Mya) facilitated global expansions of the effective population sizes of most mammalian lineages, thereby reducing the influence of genetic drift and increasing the efficiency of selection against weakly disadvantageous inserts. (Notably, the platypus, the ancestors of which may never have experienced a pronounced population-size expansion, is the only mammal for which a phase of genomic contraction is not obvious.) Although a gradual increase in N_e would be expected to influence the demography of all forms of genomic insertions, the types to respond first would be those with the most mutationally harmful effects, which would imply a ranking of average deleterious effects with pseudogenes > LTR retrotransposons > numts. Finally, if correct, the population-size expansion hypothesis should have implications for all other aspects of mammalian genome evolution. Thus, it is notable that although there has been a loss of GC content in mammalian genomes over time, the loss rate has more recently declined, possibly because of an increase in the effectiveness of biased gene conversion toward GC content (9). Because biased gene conversion operates like selection (87), with increased efficiency in larger populations, this interpretation, if correct, would provide independent support for increased N_e in post-K-T mammalian lineages.

These types of observations suggest several reasons for caution against relying on descriptions of one or a few genomes as evidence for or against the mutational-hazard hypothesis. First, owing to the fact that neutral mutations fix within $4N_e$ generations on average (46), estimates of the quantity $4N_eu$ are valid over no more than the past $4N_e$ generations, whereas the features of genomes that expand/contract in response to changes in N_eu may take much longer periods to unfold. Second, single genome sequences need not always be representative of the broader lineages from which they are derived. For example, a recent comparison of the *Daphnia pulex* genome with that from other arthropods led to the conclusion that the *Daphnia* lineage has experienced a substantial expansion in intron number (17). However, the clone selected for sequencing is known to be a member of an isolated lineage with very low recent $4N_eu$, and most of its unique introns are not even found in members of the species in other nearby populations (55). Finally, although the general syndrome of genomic bloating in response to reductions in $4N_eu$ appears to be widespread in eukaryotes, this pattern arises because of the predominance of large-scale insertions in eukaryotic genomes. The opposite pattern would be expected in species with a mutational bias toward deletion. Thus, it is notable that a syndrome of genome-size reduction in response to population size is observed in prokaryotes, which also exhibit a deletion bias (48).

Finally, it is worth emphasizing that the types of genomic alterations that accumulate in response to a reduction in $4N_e u$ need not be permanently deleterious, and may even lead to novel adaptations by modifying the possible pathways for descent with modification. For example, once established, introns allow the generation of multiple isoforms from identical precursor messenger RNAs (mRNAs) through alternative splicing, generating a level of diversity at the protein level that is otherwise not possible from a fixed number of genes, and also providing an additional potential layer of posttranscriptional gene regulation (51, 101). However, it is still unclear what proportion of alternative splicing is functional as opposed to being an indirect artifact of imperfect splicing (92), and in yeast, most introns appear to be unnecessary for normal function (93). Even pseudogenes and transposable elements can sometimes be recycled to yield functional elements. After losing their coding capacity, pseudogenes can still function as noncoding RNAs in regulating parent-gene expression (21, 103); likewise, some transposable elements have been domesticated by their host genome, where they serve as regulatory elements (11, 94). Whether this recycling of the so-called junk DNA is more the rule than the exception is still unclear, but the examples cited above illustrate the fact that evolution is an opportunistic process that can sometimes take advantage of an unfavorable situation.

PROTEIN STRUCTURE AND COMPLEX CELLULAR ADAPTATIONS

The preceding overview provides ample evidence that many aspects of gene structure respond in dramatic fashion to changes in the population-genetic environment. A key remaining question is whether the principles suggested for the proliferation of genomic modifications in response to reductions in $4N_e u$ can be extended to observations at the protein and/or cellular levels. If this were the case, then the possibility exists that the very nature of the eukaryotic cell has been molded, at least in part, by historical aspects of the rates of drift, mutation, and recombination, rather than being exclusively a product of external ecological challenges (30, 63–65, 82, 102).

Drift, Mutation, and the Modification of Protein Features

In the spirit of encouraging future research in this direction, we close with a consideration of the potential influence of drift and mutation on the evolution of various features of the protein repertoire of species. Some aspects of the mutational-hazard theory readily extend to this level. First, alternative forms of protein architecture will be underlain by gene structures that naturally impose different levels of vulnerability to deleterious mutations, and when the fitness-related consequences are sufficiently small relative to the magnitude of genetic drift, mutational pressure in the direction of more precarious protein structures will eventually drive the latter to fixation. Second, once established, modified protein structures may be exploited secondarily as substrates for novel pathways of adaptive evolution.

What are the population-genetic mechanisms by which these types of changes in protein sequences come about? The simplest way to consider how the pace of adaptive evolution might be altered by changes in N_e involves the potential fates of unconditionally beneficial single-site amino-acid alterations. Assuming an ideal random-mating population and letting u_b denote the rate of mutation to advantageous amino-acid substitutions, because the rate of input of mutations at the population level is $2N_e u_b$ and the probability of fixation of strongly beneficial mutations [defined as $4N_e s \gg 1$, where s is the selective advantage in heterozygotes (44)] is $\sim 2s$, the rate of fixation of such mutations at the population level is $\sim 4N_e u_b s$. Taken at face value, this expression suggests that the rate of adaptation will increase with population size. However, this is only strictly true if N_e and u_b are independent. Because the mutation rate declines as $\sim N_e^{-0.6}$ (66), the overall scaling of the

rate of adaptation under this simple model is actually $\sim N_e^{0.4}$, a much more gradual response to population-size increase. However, this scaling can also be misleading, as it refers to the rate of adaptation on a per-generation basis. Small organisms with large N_e have reduced generation times relative to large multicellular species with small N_e, with an approximate scaling of $N_e^{-0.8}$ (68), which implies a rate of adaptation on an absolute timescale roughly proportional to $N_e^{1.2}$. Given that the window of beneficial mutations subject to positive selection ($4N_e s \gg 1$) must increase with N_e, the opportunity for adaptive evolution on an absolute timescale via mutations with additive effects clearly increases with the effective population size.

In contrast, letting u_d denote the rate of production of deleterious mutations having neutral population dynamics ($|4N_e s| \ll 1$), $2N_e u_d$ such mutations enter the population per generation and fix with probability $1/(2N_e)$, yielding an overall rate of fixation of very mildly deleterious mutations equal to u_d per generation. We are then left with the issue of how u_d scales with population size. Considering the negative scaling of the overall mutation rate with population size, $N_e^{-0.6}$, along with the conversion to absolute time (using $N_e^{0.8}$), the scaling of the absolute-time rate of accumulation of effectively neutral but deleterious mutations, $N_e^{0.2}$, is nearly independent of N_e. However, these considerations do not take into account the fact that the distribution of deleterious mutations is strongly biased toward very small effects (4, 23, 58), which will cause a rapid increase in the pool of effectively neutral (but absolutely deleterious) mutations as N_e declines. Thus, there seems little question that the rate of fixation of mildly deleterious mutations increases with population-size reductions.

The evidence that the routes taken in protein evolution are influenced by the magnitude of random genetic drift is manifest. For example, bacteria that have relinquished a free-living form in favor of an endosymbiotic or pathogenic lifestyle occupy an environment that encourages reductions in N_e, and this is often reflected in losses of adaptive codon bias, increased accumulation of mildly deleterious amino-acid substitutions, elevated expression of molecular chaperones to accommodate protein-folding defects, and (in some cases) massive release of mobile-element activity and inactivation of entire genes (8, 12, 24, 84, 85, 107, 111). Species that have abandoned sexual reproduction and consequently experience elevated levels of genetic draft accumulate an excess of mildly deleterious amino-acid substitutions (5, 40, 89, 91). Similar observations have been made with respect to island species experiencing increased levels of drift due to population-size reductions (39, 52, 114), and even isolated populations of free-living bacteria may be vulnerable to excess deleterious-mutation accumulation (22). Finally, non-recombining organelle genomes exhibit elevated rates of accumulation of mildly deleterious amino-acid alterations (59, 60, 71), as do non-recombining sex chromosomes (2, 3, 42, 80).

The Emergence of Protein Complexes

Mutations with mild enough deleterious effects to drift to fixation are unlikely to qualitatively alter the functional core of a protein and instead may have smaller, superficial effects. For example, although optimally constructed proteins typically fold into forms that protect backbone hydrogen bonds from interactions with surrounding water molecules, many kinds of amino-acid substitutions can compromise such features, resulting in adhesive surface patches (26). A comprehensive study of a set of >100 protein orthologs with known structures revealed a gradient in the average level of exposure of backbone hydrogen bonds, with an ~50% increase from prokaryotes to unicellular eukaryotes to invertebrates and land plants to vertebrates (25). This is the same ordering noted above for the expansion of genome size and gene-structural complexity, providing suggestive evidence that the shift in population-genetic environment across these domains of life has repercussions that extend to the protein level.

Although excessive deleterious-mutation accumulation can lead to population extinction, by altering the structure of proteins, mildly detrimental mutations of this sort may also provide an opportunity for the emergence of compensatory mutations that alleviate such effects and may even open up previously inaccessible paths to adaptive evolution. Most notably, the widespread expansion of the average protein adhesiveness with a reduction in N_e suggests the possibility that the shift toward higher levels of drift in eukaryotes, particularly in multicellular species, indirectly promotes an intracellular environment conducive to the evolution of protein-protein interactions. For example, although the overall consequences of mild surface defects can be mitigated by compensatory mutations at the same locus, an alternative scenario is the development of a consortium with another potentially interacting protein in a way that hides the surface defect(s) of one or both members—e.g., modification of a monomeric protein into one that can participate in a multimeric assemblage.

Protein complexes are extraordinarily common. Of the thousands of proteins whose structures have been recorded in the PDB (Protein Data Bank), ~40% are thought to function as multimers (**Figure 3**). A number of biases may present themselves in such a survey, but taken at face value the data suggest that, relative to homomers (where different subunits are derived from the same locus), heteromers (subunits derived from distinct loci) are nearly three times more common in vertebrates than in unicellular microbes. Many potential advantages of protein-complex formation can be envisioned, including increased structural size and diversity, reduced problems of folding single large proteins, and increased flexibility for allosteric regulation and protein activation (81). However, if protein complexes are generally advantageous, we must explain (*a*) why multimers are approximately equally frequent across all domains of life, and (*b*) why heteromers are so much more abundant in taxa that are expected to be less efficient at promoting positively selected mutations.

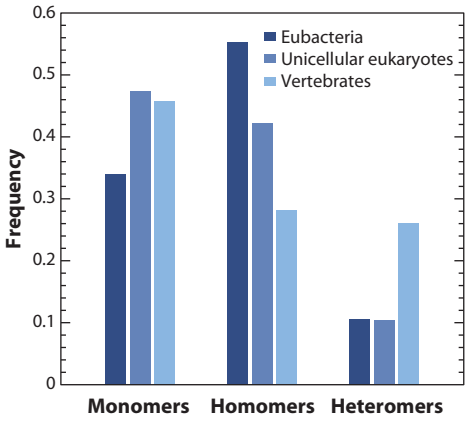

Figure 3

The distribution of protein-complex types into three broad taxonomic assemblages. Relative frequencies of the three main classes of protein structures were obtained from data deposited in 3D Complex.org v2.0 (53, 54), based primarily on taxa for which at least 20 records were available. Sample sizes are 3,545, 628, and 4,125 for eubacteria, unicellular eukaryotes, and vertebrates, respectively.

A potentially relevant point is that proteins with oligomerization potential can also be costly, providing the substrate for well-known human disorders involving inappropriate protein aggregates (e.g., Alzheimer's and Parkinson's diseases) (16), and more generally encouraging deleterious promiscuous protein-protein interactions when overexpressed (99, 104). This fine line between adaptive complexity and pathology encourages the hypothesis that the evolutionary roots of numerous protein-protein interactions may simply reside in their initial roles as compensating mechanisms for mild structural defects of monomeric proteins (25).

As in the case of the expansion of gene-structural complexity via the passive accumulation of introns, elongation of untranslated regions, etc., once a protein complex becomes established by nonadaptive mechanisms, the novel architecture may serve as the substrate for the origin of more complex cellular adaptations. In accordance with this view, substantial evidence suggests that homodimeric forms often serve as launching pads for the

evolutionary transition to more complex multimeric architectures via gene duplication and divergence of subunits (37, 95, 96). Specific examples of particularly complex traits include the flagellum (56), the nuclear pore complex (1), the spliceosome (98), the proteasome (36), and nucleosomes (79).

A precondition for the coevolutionary origin and refinement of components of a molecular interaction is colocalization of the interacting partners (49). Combined with subcellular localization of mRNAs and proteins, enhanced protein adhesiveness provides a simple starting point for such interaction. However, the opportunity for an advantageous molecular interaction need not be sufficient for its evolutionary advancement. For example, the intermediate states between one end point and another can be maladaptive, recombination can sometimes break up adaptive combinations of mutations more rapidly than they are advanced by selection, etc. Thus, to understand the conditions under which complex adaptations (requiring more than one mutation to elicit an improvement) are most likely to be promoted, we must turn to recent advances in population-genetic theory (68, 70, 108, 109). Although the work in this area is quite technical, some generalities of relevance to the current discussion can be made.

Consider the situation in which two mutations are required for the emergence of a stable complex interface, with each single-step mutation being deleterious. For example, a pair of amino-acid residues might need to be modified to produce an Arg-Asp or Glu-Lys ionic-pair or salt-bridge interaction. If N_e is small, the population will evolve via sequential fixation, with the fixation of the initially deleterious state occurring before the secondary mutation appears. However, if the population size is sufficiently large that more than one single-site mutation arises per generation, the single-site (maladaptive) mutations will always be present at a selection-mutation balance frequency of $\sim u_d/\delta$, where u_d is the total mutation rate to first-step alleles and δ is the selective disadvantage of such alleles in heterozygotes. The effective number of such alleles in a diploid population at any time is then $\sim 2N_e u_d/\delta$. Because each such allele provides the potential substrate for the origin of an adaptive second-step mutation, which upon appearance will then have a probability of fixation of $\sim 2s$, the rate of establishment of the two-site adaptation is $\sim 4N_e u_b u_d s/\delta$ per generation, where u_b is the rate of mutation of a first-step allele to an adaptive two-step allele. Again given the negative scaling of mutation rates and N_e, this quantity is expected to scale as $\sim N_e^{-0.2}$ on a per-generation basis but as $\sim N_e^{0.8}$ on an absolute-time basis, implying a substantial increase in the rate of establishment of such adaptations with increasing effective population size. In fact, for populations large enough that $4N_e u_b u_d s/\delta$ approaches 1.0, the rate of origin of haplotypes containing the adaptive complex is no longer limiting, and the time to establishment is essentially determined by the time to fixation (subsequent to origin), which is nearly independent of population size. This type of scaling with N_e does not change much with increasing complexity (i.e., increasing numbers of deleterious intermediate states) (70). Thus, population-genetic theory suggests that homomeric structures whose evolution involves intermediate deleterious states are at least as likely to arise in microbes as in multicellular species, a pattern that is in rough accordance with the similar incidence of multimers across diverse domains of life (**Figure 3**).

For diploid species, a key assumption underlying the scenario outlined above is that the double-mutant allele is immediately advantageous, or at least not detrimental, when heterozygous with alternative allelic forms. Should the allelic product of the double mutant form harmful complexes with products of ancestral alleles, fixation would be strongly inhibited because such alleles would not experience a net advantage until a sufficiently high frequency was reached that the benefits of homozygotes outweighed the disadvantages of heterozygotes. Thus, heterozygote disadvantage imposes a very strong barrier to fixation of alleles, even when there is strong homozygous

advantage, and the larger the population size, the lower the likelihood that such alleles will stochastically drift to a high enough frequency to vault the selection barrier.

As an example of why such issues are important, consider the domain-swapping model of Bennett et al. (10), whereby a monomeric protein is predisposed to making an evolutionary transition to a homodimeric structure by virtue of its initial structure—e.g., a monomeric subunit that folds to include an interface between two domains separated by a loop. A single large deletion within the loop would prevent the formation of an interface within a single monomeric subunit, while still allowing the swapping of complementary domains across two subunits to yield a homodimer. If, however, the short-looped protein competed with the internal binding of the long-looped version (or vice versa), harmful intermediates would be produced in heterozygotes. Although a number of plausible cases of domain-swapped homodimers have been discovered (57), the preceding arguments suggest that the establishment of such forms is most likely in populations experiencing a high magnitude of random genetic drift, or in haploid species that never experience heterozygosity.

In principle, heteromers may harbor much more potential for long-term evolutionary diversification than homomers, as the amino-acid sequences of different subunits are only free to diverge substantially when they are encoded by different loci. However, although heteromers may ultimately be required for the construction of complex cellular features, their de novo emergence from preexisting loci may be possible in only a narrow set of population-genetic environments. Perhaps the most critical issue is whether the loci involved recombine. If the loci are effectively completely linked, all of the preceding arguments apply, as the consortium of genes will behave like a single locus. However, recombination (either crossing-over within chromosomes or random segregation of different chromosomes) between loci can impose a strong barrier to the emergence of a two-locus adaptation (68, 110).

Consider again the case in which two mutations are required for a novel adaptation, one at each of two loci, with the intermediate (single-mutant) haplotypes being deleterious. Each of the single-locus mutations will then be kept at low frequency by selection-mutation balance until there is an opportunity for expansion of a double mutant. Although recombination between the alternative single-mutant haplotypes can enhance the rate of production of the adaptive combination, because the nonmutant (ancestral) haplotype will predominate in the population at the time of first appearance of the adaptive combination, almost all subsequent recombination events will convert the double mutant back to the maladapted single-mutation types. Thus, if the rate of recombination between loci exceeds the selective advantage of the double mutant, it will be essentially impossible for the adaptive complex to become established unless the power of drift is sufficiently large to allow the intermediate-state haplotypes to simply drift to high enough frequencies to overcome the recombinational barrier. This again shows that despite their potential advantages when fixed in a population, some types of protein assemblages can be strongly inhibited from establishing in populations with large N_e, especially if the intermediate states are deleterious.

One final point addresses the issue of heteromeric complexes in a way that directly links to the mutational-hazard hypothesis. Should the function of a protein complex be equally accomplishable via a homomer or heteromer, then aside from the requirement for a gene duplication, development of the latter will be inhibited owing to the fact that heterodimers present double the target size for inactivating mutations. This excess cost of mutation is expected to be small, as it is equivalent to the per-locus mutation rate to defective alleles, which is sufficient only to prevent heterodimer establishment if it exceeds the power of random genetic drift (76). However, because the mutational burden of an entire gene locus is greater than that for small embellishments to a gene, as in the case of the expansion of

gene-structural complexity, the mutational burden of excess protein complexity is likely to be sufficient to inhibit the emergence of many forms of protein-protein complexes in populations of unicellular species with very large N_e.

SUMMARY POINTS

1. Although it is commonly assumed that virtually all aspects of biodiversity, including those at the genomic and subcellular levels, reflect long-term adaptive tuning to persistent ecological challenges, there is substantial empirical evidence that the three major nonadaptive forces of evolution—mutation, recombination, and random genetic drift—vary by orders of magnitude among phylogenetic lineages.
2. Recent investigations in theoretical population genetics suggest that this phylogenetic range of the population-genetic environment is sufficient to impose major differences in the pathways that are open or closed to evolutionary exploitation in microbes versus multicellular eukaryotes.
3. The mutational-hazard theory, which postulates that virtually all forms of excess DNA impose a weak mutational burden, provides a null hypothesis for interpreting patterns of genome evolution, yielding a potentially unifying explanation for the dramatic gradient in genome size and many aspects of gene-structural complexity across the tree of life. These patterns are not readily explained by alternative hypotheses that invoke DNA as an architectural feature of the cell or that postulate genomic streamlining as a mechanism for enhancing replication rates or minimizing problems with nutrient limitation.
4. Ample evidence also exists that reductions in the efficiency of selection, associated with elevated rates of random genetic drift in various eukaryotic lineages, have been sufficient to facilitate the accumulation of mildly deleterious amino-acid substitutions in proteins as well as to secondarily encourage the emergence of novel protein-protein complexes.
5. The overall interpretation of these results is that a complete understanding of patterns of variation at both the genomic and proteomic levels, and perhaps beyond, cannot be achieved with a framework that assumes natural selection to be the only relevant evolutionary force and fails to consider the unique combinations of mutation, recombination, and random genetic drift experienced by various taxa.

FUTURE ISSUES

1. As the principles outlined above pertain to all levels of organismal diversity, in the bottom-up approach to understanding the mechanisms of evolution, it is now time to move beyond the level of genome architecture to higher-order features such as protein structure. This will provide a natural bridge between genomic and phenotypic evolution, and hence a potential entrée into the evolution of cellular features.
2. Most previous studies of protein evolution have focused on primary sequence structure, but compelling evidence now exists that the transitions from prokaryotes to unicellular eukaryotes to multicellular species have been accompanied by changes in the general architectural features of proteins, including the evolution of multimeric structures.

3. Virtually all features of the cell are derived from protein assemblages, and it is now known that many of these originated from ancestral homomeric cores around which other components were recruited by gene duplication and divergence. Thus, general insights into how novel cellular attributes evolve are achievable if a research agenda focused on the comparative architecture of orthologous proteins across divergent taxa can be combined with population-genetic theory on the evolution of complex adaptations.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work has been supported by grants to M.L. from the National Science Foundation (EF-0328516 and EF-0827411), National Institutes of Health (R01-GM036827), and U.S. Department of the Army (ONRBAA10-002). F.C. is supported by a Marie Curie International Incoming Fellowship (254202).

LITERATURE CITED

1. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, et al. 2007. The molecular architecture of the nuclear pore complex. *Nature* 450:695–701
2. Bachtrog D. 2006. Expression profile of a degenerating neo-Y chromosome in *Drosophila*. *Curr. Biol.* 16:1694–99
3. Bachtrog D. 2008. The temporal dynamics of processes underlying Y chromosome degeneration. *Genetics* 179:1513–25
4. Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat. Rev. Genet.* 8:619–31
5. Barraclough TG, Fontaneto D, Ricci C, Herniou EA. 2007. Evidence for inefficient selection against deleterious mutations in cytochrome oxidase I of asexual bdelloid rotifers. *Mol. Biol. Evol.* 24:1952–62
6. Barton NH. 2010. Genetic linkage and natural selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365:2559–69
7. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, et al. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327:836–40
8. Belda E, Moya A, Bentley S, Silva FJ. 2010. Mobile genetic element proliferation and gene inactivation impact over the genome structure and metabolic capabilities of *Sodalis glossinidius*, the secondary endosymbiont of tsetse flies. *BMC Genomics* 11:449
9. Belle EM, Duret L, Galtier N, Eyre-Walker A. 2004. The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. *J. Mol. Evol.* 58:653–60
10. Bennett MJ, Choe S, Eisenberg D. 1994. Domain swapping: entangling alliances between proteins. *Proc. Natl. Acad. Sci. USA* 91:3127–31
11. Biemont C, Vieira C. 2006. Genetics: junk DNA as an evolutionary force. *Nature* 443:521–24
12. Burke GR, Moran NA. 2011. Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. *Genome Biol. Evol.* 3:195–208
13. Cavalier-Smith T. 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J. Cell Sci.* 34:247–78
14. Cavalier-Smith T. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann. Bot.* 95:147–75

15. Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* 10:195–205
16. Chiti F, Dobson CM. 2009. Amyloid formation by globular proteins under native conditions. *Nat. Chem. Biol.* 5:15–22
17. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, et al. 2011. The ecoresponsive genome of *Daphnia pulex*. *Science* 331:555–61
18. Dawson KJ. 1999. The dynamics of infinitesimally rare alleles, applied to the evolution of mutation rates and the expression of deleterious mutations. *Theor. Popul. Biol.* 55:1–22
19. Dumont BL, Payseur BA. 2008. Evolution of the genomic rate of recombination in mammals. *Evolution* 62:276–94
20. Dumont BL, Payseur BA. 2011. Evolution of the genomic recombination rate in murid rodents. *Genetics* 187:643–57
21. Duret L, Chureau C, Samain S, Weissenbach J, Avner P. 2006. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* 312:1653–55
22. Escobar-Paramo P, Ghosh S, DiRuggiero J. 2005. Evidence for genetic drift in the diversification of a geographically isolated population of the hyperthermophilic archaeon *Pyrococcus*. *Mol. Biol. Evol.* 22:2297–303
23. Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8:610–18
24. Fares MA, Moya A, Barrio E. 2004. GroEL and the maintenance of bacterial endosymbiosis. *Trends Genet.* 20:413–16
25. Fernández A, Lynch M. 2011. Non-adaptive origins of interactome complexity. *Nature* 474:502–5
26. Fernández A, Zhang X, Chen J. 2008. Folding and wrapping soluble proteins exploring the molecular basis of cooperativity and aggregation. *Prog. Mol. Biol. Transl. Sci.* 83:53–87
27. Froula JL, Francino MP. 2007. Selection against spurious promoter motifs correlates with translational efficiency across bacteria. *PLoS One* 2:e745
28. Gillespie JH. 2000. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* 155:909–19
29. Gordo I, Charlesworth B. 2000. The degeneration of asexual haploid populations and the speed of Muller's ratchet. *Genetics* 154:1379–87
30. Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF. 2010. Cell biology: irremediable complexity? *Science* 330:920–21
31. Gregory TR, Witt JD. 2008. Population size and genome size in fishes: a closer look. *Genome* 51:309–13
32. Hahn MW, Stajich JE, Wray GA. 2003. The effects of selection against spurious transcription factor binding sites. *Mol. Biol. Evol.* 20:901–6
33. Hartfield M, Otto SP, Keightley PD. 2010. The role of advantageous mutations in enhancing the evolution of a recombination modifier. *Genetics* 184:1153–64
34. Hessen DO, Jeyasingh PD, Neiman M, Weider LJ. 2010. Genome streamlining and the elemental costs of growth. *Trends Ecol. Evol.* 25:75–80
35. Hessen DO, Jeyasingh PD, Neiman M, Weider LJ. 2010. Genome streamlining in prokaryotes versus eukaryotes. *Trends Ecol. Evol.* 25:320–21
36. Hughes AL. 1997. Evolution of the proteasome components. *Immunogenetics* 46:82–92
37. Ispolatov I, Yuryev A, Mazo I, Maslov S. 2005. Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Res.* 33:3629–35
38. Jeffreys AJ, Neumann R. 2009. The rise and fall of a human recombination hot spot. *Nat. Genet.* 41:625–29
39. Johnson KP, Seger J. 2001. Elevated rates of nonsynonymous substitution in island birds. *Mol. Biol. Evol.* 18:874–81
40. Johnson SG, Howard RS. 2007. Contrasting patterns of synonymous and nonsynonymous sequence evolution in asexual and sexual freshwater snail lineages. *Evolution* 61:2728–35
41. Johnson T. 1999. The approach to mutation-selection balance in an infinite asexual population, and the evolution of mutation rates. *Proc. Biol. Sci.* 266:2389–97

42. Kaiser VB, Charlesworth B. 2010. Muller's ratchet and the degeneration of the *Drosophila miranda* neo-Y chromosome. *Genetics* 185:339–48
43. Keightley PD, Otto SP. 2006. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature* 443:89–92
44. Kimura M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 47:713–19
45. Kimura M. 1967. On the evolutionary adjustment of spontaneous mutation rates. *Genet. Res.* 9:23–34
46. Kimura M, Ohta T. 1969. The average number of generations until fixation of a mutant gene in a finite population. *Genetics* 61:763–71
47. Koonin EV. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol.* 11:209
48. Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* 19:1450–54
49. Kuriyan J, Eisenberg D. 2007. The origin of protein interactions and allostery in colocalization. *Nature* 450:983–90
50. Lane N, Martin W. 2010. The energetics of genome complexity. *Nature* 467:929–34
51. Lareau LF, Green RE, Bhatnagar RS, Brenner SE. 2004. The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol.* 14:273–82
52. Legrand D, Tenaillon MI, Matyot P, Gerlach J, Lachaise D, Cariou ML. 2009. Species-wide genetic variation and demographic history of *Drosophila sechellia*, a species lacking population structure. *Genetics* 182:1197–206
53. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA. 2008. Assembly reflects evolution of protein complexes. *Nature* 453:1262–65
54. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA. 2006. 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* 2:e155
55. Li W, Tucker AE, Sung W, Thomas WK, Lynch M. 2009. Extensive, recent intron gains in *Daphnia* populations. *Science* 326:1260–62
56. Liu R, Ochman H. 2007. Stepwise formation of the bacterial flagellar system. *Proc. Natl. Acad. Sci. USA* 104:7116–21
57. Liu Y, Eisenberg D. 2002. 3D domain swapping: as domains continue to swap. *Protein Sci.* 11:1285–99
58. Loewe L, Charlesworth B. 2006. Inferring the distribution of mutational effects on fitness in *Drosophila*. *Biol. Lett.* 2:426–30
59. Lynch M. 1996. Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. *Mol. Biol. Evol.* 13:209–20
60. Lynch M. 1997. Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. *Mol. Biol. Evol.* 14:914–25
61. Lynch M. 2002. Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. USA* 99:6118–23
62. Lynch M. 2006. The origins of eukaryotic gene structure. *Mol. Biol. Evol.* 23:450–68
63. Lynch M. 2007. The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet.* 8:803–13
64. Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci. USA* 104(Suppl. 1):8597–604
65. Lynch M. 2007. *The Origins of Genome Architecture*. Sunderland, MA: Sinauer
66. Lynch M. 2010. Evolution of the mutation rate. *Trends Genet.* 26:345–52
67. Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* 107:961–68
68. Lynch M. 2010. Scaling expectations for the time to establishment of complex adaptations. *Proc. Natl. Acad. Sci. USA* 107:16577–82
69. Lynch M. 2011. The lower bound to the evolution of mutation rates. *Gen. Biol. Evol.* In press
70. Lynch M, Abegg A. 2010. The rate of establishment of complex adaptations. *Mol. Biol. Evol.* 27:1404–14
71. Lynch M, Blanchard JL. 1998. Deleterious mutation accumulation in organelle genomes. *Genetica* 102–103:29–39
72. Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302:1401–4
73. Lynch M, Conery JS, Bürger R. 1995. Mutation accumulation and the extinction of small populations. *Am. Nat.* 146:489–518

74. Lynch M, Conery JS, Bürger R. 1995. Mutational meltdowns in sexual populations. *Evolution* 49:1067–80
75. Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. *Science* 311:1727–30
76. Lynch M, O'Hely M, Walsh B, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* 159:1789–804
77. Lynch M, Scofield DG, Hong X. 2005. The evolution of transcription-initiation sites. *Mol. Biol. Evol.* 22:1137–46
78. Lynch M, Walsh B. 1998. *Genetics and Analysis of Quantitative Traits.* Sunderland, MA: Sinauer
79. Malik HS, Henikoff S. 2003. Phylogenomics of the nucleosome. *Nat. Struct. Biol.* 10:882–91
80. Marais GA, Nicolas M, Bergero R, Chambrier P, Kejnovsky E, et al. 2008. Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia*. *Curr. Biol.* 18:545–49
81. Marianayagam NJ, Sunde M, Matthews JM. 2004. The power of two: protein dimerization in biology. *Trends Biochem. Sci.* 29:618–25
82. Martin W. 2010. Evolutionary origins of metabolic compartmentalization in eukaryotes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365:847–55
83. Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* 440:41–45
84. Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. USA* 93:2873–78
85. Moran NA, Plague GR. 2004. Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.* 14:627–33
86. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, et al. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327:876–79
87. Nagylaki T. 1983. Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. USA* 80:6278–81
88. Nei M, Tajima F. 1981. Genetic drift and estimation of effective population size. *Genetics* 98:625–40
89. Neiman M, Hehman G, Miller JT, Logsdon JM Jr, Taylor DR. 2010. Accelerated mutation accumulation in asexual lineages of a freshwater snail. *Mol. Biol. Evol.* 27:954–63
90. Paigen K, Petkov P. 2010. Mammalian recombination hot spots: properties, control and evolution. *Nat. Rev. Genet.* 11:221–33
91. Paland S, Lynch M. 2006. Transitions to asexuality result in excess amino acid substitutions. *Science* 311:990–92
92. Pan Q, Saltzman AL, Kim YK, Misquitta C, Shai O, et al. 2006. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.* 20:153–58
93. Parenteau J, Durand M, Veronneau S, Lacombe AA, Morin G, et al. 2008. Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Mol. Biol. Cell* 19:1932–41
94. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, et al. 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* 7:597–606
95. Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA. 2007. Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.* 8:R51
96. Reid AJ, Ranea JA, Orengo CA. 2010. Comparative evolutionary analysis of protein complexes in *E. coli* and yeast. *BMC Genomics* 11:79
97. Rho M, Zhou M, Gao X, Kim S, Tang H, Lynch M. 2009. Independent mammalian genome contractions following the KT boundary. *Genome Biol. Evol.* 1:2–12
98. Scofield DG, Lynch M. 2008. Evolutionary diversification of the Sm family of RNA-associated proteins. *Mol. Biol. Evol.* 25:2255–67
99. Semple JI, Vavouri T, Lehner B. 2008. A simple principle concerning the robustness of protein complex activity to changes in gene expression. *BMC Syst. Biol.* 2:1
100. Speijer D. 2011. Does constructive neutral evolution play an important role in the origin of cellular complexity? Making sense of the origins and uses of biological complexity. *Bioessays* 33:344–49
101. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, et al. 2005. Function of alternative splicing. *Gene* 344:1–20

102. Stoltzfus A. 1999. On the possibility of constructive neutral evolution. *J. Mol. Evol.* 49:169–81
103. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 453:534–38
104. Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B. 2009. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* 138:198–208
105. Vieira-Silva S, Touchon M, Rocha EP. 2010. No evidence for elemental-based streamlining of prokaryotic genomes. *Trends Ecol. Evol.* 25:319–20
106. Wang J. 2005. Estimation of effective population sizes from data on genetic markers. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360:1395–409
107. Warnecke T, Rocha EP. 2011. Function-specific accelerations in rates of sequence evolution suggest predictable epistatic responses to reduced effective population size. *Mol. Biol. Evol.* In press
108. Weinreich DM, Chao L. 2005. Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution* 59:1175–82
109. Weissman DB, Desai MM, Fisher DS, Feldman MW. 2009. The rate at which asexual populations cross fitness valleys. *Theor. Popul. Biol.* 75:286–300
110. Weissman DB, Feldman MW, Fisher DS. 2010. The rate of fitness-valley crossing in sexual populations. *Genetics* 186:1389–410
111. Wernegreen JJ, Moran NA. 1999. Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. *Mol. Biol. Evol.* 16:83–97
112. Whitney KD, Garland T Jr. 2010. Did genetic drift drive increases in genome complexity? *PLoS Genet.* 6:e1001080
113. Wible JR, Rougier GW, Novacek MJ, Asher RJ. 2007. Cretaceous eutherians and Laurasian origin for placental mammals near the K/T boundary. *Nature* 447:1003–6
114. Woolfit M, Bromham L. 2005. Population size and molecular evolution on islands. *Proc. Biol. Sci.* 272:2277–82

Annual Review of Genomics and Human Genetics

Volume 12, 2011

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* articles may be found at http://genom.annualreviews.org/errata.shtml